Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-020-00589-6 OPEN

Expressivism, self-knowledge, and rational agency

Stephen Blackwood [1]✉

One family of thought about self-knowledge has argued that authoritative self-ascriptions express a form of higher-order knowledge whose special character is explained by the role that knowledge plays in rational agency. In contrast to this "regulative model", according to Wittgenstein's treatment of self-knowledge authoritative self-ascription of one's present-tense mental states is explained by the fact that sincere self-ascriptions express the very states they self-ascribe. The Wittgensteinian account is epistemologically deflationary, and it makes no use of higher-order thought to account for the distinctive features of self-ascriptions. It is argued that the regulative model faces difficulties that both undermine it and reinforce the Wittgensteinian explanation. Making use of ideas from Donald Davidson and Richard Moran, an alternative first-order sketch of rational agency consistent with the expressivist view is offered.

[1] Grenfell Campus, Memorial University of Newfoundland, Corner Brook, NL A1C 5S7, Canada. ✉email: sblackwood@grenfell.mun.ca

Introduction

It can be difficult to see how there might be room for an account of either self-knowledge or of rationality that does not make use of a second-level of mind—a level of second-order judgments that, under appropriate conditions, can qualify as knowledge of our first-order mental states, and a level of second-order critical judgments that monitor and regulate those first-order states, preserving consistency and coherence among them. Indeed, from one perspective self-knowledge and rationality will seem to be intrinsically linked: we can maintain rational order among our beliefs, desires, and intentions only if we have knowledge of them. One school of thought about self-knowledge, which I will call the *regulative model*, argues that self-ascriptions of mental states express a species of knowledge whose special character is explained by the role it plays in rationality. By contrast, a Wittgensteinian treatment of the special character of self-knowledge is explained by the fact that sincere self-ascriptions express the states they self-ascribe. The Wittgensteinian account is epistemologically deflationary, and it makes no use of higher-order thoughts to account for the distinctive features of self-ascriptions. I argue that the regulative model faces difficulties that both undermine it and reinforce the Wittgensteinian explanation. Drawing on ideas from Donald Davidson and Richard Moran, I then describe a perspective on rationality that makes no use of a second-level of mind. It is not difficult to see in the regulative model a residue of the Cartesian theater of mind, with the players on stage under the gaze of audience and critic. But for all its initial attractiveness, the regulative model of mind is problematic; it will help to see that it is not mandatory.

The problem of self-knowledge: self-asymmetries and other-asymmetries

Our first-person present tense ascriptions of contentful mental states (for example, of beliefs, desires, and intentions), and phenomenal states (such as pains and the like) are thought to differ in significant ways from our ascriptions of those states to others. For example, when a person ascribes a mental state that *p* to another, she must do so on the evidence provided by the utterances and actions of the other. But unlike other-person ascriptions, self-ascriptions are typically groundless, or immediate: demands that we justify our self-ascriptions, or explain how we know we have the mental states we self-ascribe, are generally deemed inappropriate. Furthermore, assuming their sincerity, self-ascriptions that are not made on the basis of behavioral evidence are highly likely to be correct. This likelihood of correctness does not extend to ascriptions of mental states to others. Thus, persons appear to possess a level of authority in their self-ascriptions that, while it falls short of infallibility, is far greater than they enjoy in their attributions to others.

Those who accept that these asymmetries obtain follow one of two broad explanatory paths. In recent years, an epistemically deflationary approach has gained currency. According to such views, the authority and immediacy granted to typical self-ascriptions are not to be explained in terms of any privileged perceptual position the subject occupies with respect to her own mental states, nor in any advantage in the amount or quality of evidence she might have for them. Rather, it is based on some *non-epistemic* feature of self-ascriptions. Still, for most philosophers the question remains an epistemic one. The task as they see it is to say how to incorporate the asymmetries into an account that shows how our self-ascriptions qualify as expressions of knowledge, that is, as warranted true beliefs about our mental states.

In what follows I will first look at a selection of philosophers who try to explain these ascriptive asymmetries by drawing an essential connection between our capacity for self-knowledge and our status as critical reasoners and rational agents. I then briefly describe a deflationary (Wittgensteinian) expressivist explanation of the asymmetries. Finally, I argue that the former accounts suffer from problems that both undermine it and encourage the expressivist account of self-ascriptions.

The challenge to an epistemic treatment of self-knowledge

According to Paul Boghossian (1998), the challenge for epistemological treatments of the distinctive character of our self-ascriptions is this: On the one hand, the idea of self-knowledge—the capacity to formulate justified true beliefs about our mental states—is presupposed by many of the concepts (for example, intentional action) that are fundamental to our ordinary self-conception. Consequently, insofar as we cannot see our way to an alternative self-conception, a skeptical view that denies such a capacity must be rejected. On the other hand, upon inspection we find that the options for an epistemic account come up wanting. The conclusion is that, while we cannot do without the idea of self-knowledge, we have little idea what form an epistemic explanation of self-knowledge might take.

Boghossian arrives at this conclusion by identifying apparently irreconcilable features of self-knowledge. He asks how we might account for our capacity to produce justified true beliefs about contentful thoughts, beliefs, or fears. For example, immediately upon thinking 'Even great composers write lousy arias,' one knows what one has thought (Boghossian, 1998, p. 152). As Boghossian sees it there are three possible avenues for an explanation of such knowledge to take. One could show how such judgments are derived from (1) inference, (2) some inner analog of observation, or (3) some other *non-empirical* basis (Boghossian, 1998, pp. 149–150).

The inferential option seems hopeless, since it denies, rather than explains, the apparently immediate or groundless character of our self-attributions. But even worse, for many self-ascriptions the type of behavioral evidence to which an inferential account would need to appeal is not available to the subject. Sitting quietly at my desk I might think 'Even great philosophers sometimes make mistakes', knowing full well what I have thought despite lacking any behavioral evidence that might serve in premises for an inference to a self-ascription of the thought.

Boghossian also argues that an internalist conception of justification, to which many philosophers remain sympathetic, demands that self-knowledge be non-inferential. On the internalist view of justification, one is justified in the second-order belief that one believes that *p* only if one recognizes that one has the further belief upon which that belief rests (say, a belief that q). So internalism about justification requires that I already know that I have the beliefs that serve as premises in any inference that supports my self-ascription, and that requirement sends us off on a vicious regress. We are left to conclude that there must be a way of knowing the contents of our mental states (including thoughts) non-inferentially. But, then, either self-knowledge is based on inner observation, or it is grounded on "nothing empirical" (Boghossian, 1998, p. 156).

The inner observation option, while perhaps not so immediately counter-intuitive, is also untenable. Given certain widely accepted externalist claims about the determinants of thought content, it follows that we could not know the content of our thoughts through mere inspection of their observable intrinsic (narrow) properties. To know that one is thinking of water, and not *twin water*, one needs to know that one's thought is caused by H_2O and not A_2Z. However, no inner observation will give us knowledge of the content determining external property.

Consequently, any judgment about what we are thinking will be susceptible to the skeptical charge that we don't know what content we are attributing to ourselves. (Boghossian, 1998, p. 166). So, in brief, Boghossian's argument goes. But if he is right, then we are left with his third option, that self-knowledge is "based on nothing". What can this mean?

Normally knowledge of a contingent proposition is grounded on observation or on an inference from observation. Such knowledge involves a "cognitive achievement", and its epistemology is always "substantial" (Boghossian, 1998, p. 165). But knowledge that is "based on nothing" does not derive from any cognitive achievement and its epistemology is therefore "insubstantial." Boghossian offers a few examples of potentially baseless, or insubstantial, knowledge. First, there are certain self-regarding indexical propositions, such as "I am here now", that are, according to him, true and justified as soon as they are entertained. Secondly, some philosophers suppose that there are self-regarding, self-verifying propositions that, on being thought, constitute one as being in the state they indicate. For example, there may be no fact of the matter about my being jealous of my friend prior to my judgment that I am, but my sincerely thinking it makes it so. In such cases, my judgments would be both true and justified, even though they are not grounded on any empirical evidence – observation, or inference from observation, would be irrelevant to the question of their truth or warrant.

A third kind of insubstantial self-knowledge claim considered by Boghossian is what Tyler Burge (1998) terms "basic self-knowledge"—self-ascriptions of the form 'I am thinking that *p*'. Burge argues that in thinking such second-order thoughts one also thinks the first-order thoughts they are about. Their self-referential, logically self-verifying character ensures that such thoughts are both true and warranted. Since the second-order thought takes its content from the first-order thought it "contains", it is assured to have whatever content the first-order thought has. Thus, the account evades the threat posed by externalist theories of content. Boghossian does not disagree; however, he observes that such "basic" cases are not the usual case, and the account has nothing to say about propositional attitudes in general.[1]

The limited range of cases covered by the three accounts just considered points to the difficulty one faces in arriving at an epistemologically insubstantial explanation of the authority we are said to enjoy with regard to our thoughts generally. However, for Boghossian the lack of wider application is not the most pressing issue such accounts face. The main problem is that the truth of second-order judgments in the three sorts of cases is *guaranteed*. But this is not in keeping with our ordinary conception of self-knowledge—first-person authority is not thought to equal infallibility. With respect to the limitations to authority, Boghossian writes: "I know of no convincing alternative to the following type of explanation: the difference between getting it right and failing to do so (either through ignorance or through error) is the difference between being in an epistemically favorable position with relevant evidence—and not" (1998, p. 167). If so, it would appear that we must make room for "genuine cognitive achievement" in our account of self-knowledge after all, for otherwise we will have no way of making sense of our admitted failures of self-knowledge

It seems that we are in a quandary. The foregoing offers persuasive reasons to reject explanations of immediate and authoritative self-ascriptions in terms of observational or inferential "cognitive achievement". Unfortunately, the alternative explanatory approach—that our authority is "based on nothing"—is also deeply problematic due to its limited range of application and inability to account for our admitted fallibility. Indeed, according to Boghossian the recognition of our fallibility suggests that successful self-attribution must involve some sort of epistemological achievement. This returns us to the idea that authoritative self-ascription about our mental lives must, after all, be based on some form of cognitive achievement, even if the possibilities canvased all fall short. This is not to say that Boghossian thinks a solution is impossible; but, he says, "we have a serious problem explaining our ability to know our own thoughts, a problem that has perhaps not been sufficiently appreciated" (1998, p. 172).

In the following sections I will examine a family of epistemological accounts of authoritative self-ascription that, while they avoid appeal to any observational or inferential basis, still argue for a form of cognitive achievement. Sydney Shoemaker, Tyler Burge, and Richard Moran have each argued for an essential link between the authority that is thought to accrue to self-ascriptions and our status as rational subjects. Each argues that an understanding of how our self-ascriptions count as knowledge is to be found in consideration of the role played by first-person second-order judgments and beliefs in rational agency. After providing challenges to those views, I will introduce a non-epistemic way of accounting for authoritative self-ascriptions that allows for rationality without the need to deploy second-order judgments.

Shoemaker: the necessity of self-awareness for rationality

In 'On Knowing One's Mind' (1996), Shoemaker contends that the rationalizing change of belief requires self-knowledge ("or at least something very much like it", as he puts it [Shoemaker, 1996, p. 31]). More specifically, it requires (1) second-order beliefs about what one's current first-order beliefs and desires are, (2) second-order desires to promote consistency in those first-order beliefs, and (3) second-order beliefs regarding what changes are required to satisfy those second-order desires (Shoemaker, 1996, p. 33). Furthermore, he offers a *reductio* argument against a phenomenon that he calls 'self-blindness' (a condition wherein one can only recognize the truth of one's second-order beliefs by interpreting one's own behavior) to show that knowledge of one's first-order mental states must be gained through a form of immediate privileged access he terms "self-acquaintance" (Shoemaker, 1996, p. 25). The argument goes like this: If self-knowledge by self-acquaintance were an optional component of our rational lives—in other words, if self-blindness were possible—then a case in which a person lacked knowledge from self-acquaintance would be revealed by discrepancies between her behavior and the behavior of one who possessed such knowledge (a "normal" person, as Shoemaker puts it). However, he argues, no such discrepancy would be found. This leaves two options: either (1) deny that we have self-knowledge by self-acquaintance, or (2) take the fact that no difference could be discerned between a self-blind and a self-acquainted person as a *reductio* of the possibility of self-blindness and, thus, as proof of the necessity of privileged self-knowledge (Shoemaker, 1996, pp. 36, 39).

Since he thinks the first option is absurd, Shoemaker concludes that second-order judgments about our beliefs and desires, bearing the marks of self-acquaintance, are required for rational deliberation. Shoemaker suggests that this is also the mechanism through which we express our agency: We are responsible for our beliefs and other mental states in virtue of the fact that we can exercise control over them through our second-order deliberations on their rational standing (Shoemaker, 1996, p. 28). Given that this requires knowledge of what those states are, it follows that self-knowledge (by self-acquaintance) is essential to our status as rational agents. I shall refer to accounts of self-knowledge that, like Shoemaker's, link the monitoring or regulative role of second-order beliefs to agency as rational agency models of self-knowledge. As I read him, Tyler Burge also subscribes to this general view.

Burge: self-knowledge and the requirements of critical rationality

In 'Our Entitlement to Self-Knowledge' (1998) Burge also takes the justified true second-order beliefs that are self-knowledge to be a fundamental component of critical rationality. He argues that the truth and warrant of second-order judgments constitutive of self-knowledge is connected to the entitlement we have to knowledge claims in general. This is because critical reason is an essential component of the knowledge enterprise. That said, he also argues that the kind of entitlement attached to second-order judgments must be distinct from that in ordinary perceptual belief. As he puts it, "there must be a non-contingent, rational relation between relevant first-person judgments and their subject matter or truth", a relation that is constitutive of critical reason (Burge, 1998, p. 246). More specifically, our entitlement to self-knowledge claims is tied to our status as critical reasoners, to our ability to operate in accord with norms of reason, even if these norms cannot be articulated by the reasoner.

With respect to our reflective second-order beliefs in particular, our entitlement to them derives from the role they play in ensuring the reasonability of the whole process of critical reasoning. If our judgments about our first-order mental states and their interrelations were not rational (if we lacked entitlement to them), then our reflection on those states would fail to add to the rationality of the whole reasoning process. But, Burge says, "reflection does add a rational element to the reasonability of reasoning. It gives one rational control over one's reasoning". As he goes on to say, "critical reasoning just is reasoning in which norms of reason apply to how attitudes should be affected partly on the basis of reasoning that derives from judgments about one's attitudes" (Burge, 1998, p. 249). Thus, our status as critical reasoners confers epistemic entitlement on our second-order judgments about our first-order beliefs. However, Burge adds, entitlement is not enough—for similar reasons those second-order judgments must also be generally true; otherwise the link between the two levels of belief, and consequently one's ability to reflect critically, would break down. If reflection bore on the truth of our second-order beliefs in a merely contingent way, then the reason-guiding and coherence-making functions of critical reflection would fail. Or if we were entitled to our second-order judgments but they were systematically mistaken, then we could not be critical reasoners. "For critical reasoning requires rational integration of one's higher-order evaluations with one's first-order, object-oriented reasoning. … If the two came radically apart, or were only accidentally connected, critical reasoning would not occur" (Burge, 1998, p. 250).

So for Burge, knowledgeable self-ascriptions of mental states are a basic component of critical reflection; if self-ascriptive judgments weren't reliably correct, then the critical reflection in which we engage could not get off the ground. Like Shoemaker, Burge also sees this second-order capacity as essential to agency—we can be held responsible for our mental states only because we are capable of reviewing our reasons and reasoning (Burge, 1998, p. 258).

Moran: the importance of a non-alienated first-person perspective

Shoemaker and Burge share the assumption that our authoritative self-ascriptions express second-order beliefs. Each argues that the distinctive character of the self-ascriptions taken as expressive of these second-order beliefs—their groundlessness and unparalleled security—as well as their warrant, reflect intrinsic links between self-knowledge and rational agency, where the latter is construed in terms of the rational control a subject exercises over her mental life through her second-order deliberation on it.

Richard Moran (2001) also ties a proper understanding of the nature of self-knowledge to our capacity for rational deliberation and agency. He is sympathetic to the general tenor of their views, arguing that a proper discussion must go beyond an explanation of the special mode of awareness and security characteristic of avowals: "[t]he special features of first-person awareness cannot be understood by thinking of it purely in terms of epistemic access. … Rather we must think of it in terms of the special responsibilities the person has in virtue of the mental life in question being his own" (Moran, 2001, p. 32). However, he argues that the scope of the explanations they offer is too restricted, and that they fail to fully account for the nature of what he terms "genuine" first-person awareness of one's own beliefs.

Accounting for genuine self-knowledge requires that one see one's beliefs and other attitudes as "expressive of his various and evolving relations to his environment, and not as a mere succession of representations (to which, for some reason, he is the only witness" (Moran, 2001, p. 32). As he sees it, talk of consciousness carries with it a host of implications for the subject and her responsibilities and commitments—the epistemic perspective she takes toward herself has significant consequences for her relation to herself and her self-conception. Consequently, at the center of his account is a view about the place of deliberation and the role it plays in self-constitution, in *making up* one's mind about what one ought to and will believe, desire, or intend.

Moran proceeds by distinguishing between *theoretical* and *deliberative* self-knowledge (Moran, 2001, p. 55). The former is essentially third-personal, in the sense that it is restricted in scope or perspective to the *description* of the psychological facts about oneself. Theoretical inquiry into one's states ends with a second-order belief about the content and/or quality of them. Alternatively, one can adopt what Moran calls a *deliberative stance* toward oneself (Moran, 2001, p. 59). To judge from this perspective is to undertake practical reflection, the end point of which is not merely a belief about the content or character of a first-order mental state, but a commitment to, or endorsement of, the content of that state. Such inquiry conforms to what Moran terms the *Transparency Condition*, according to which such questions as "Do I believe (desire, intend, etc.) that p" are answered by reflection on (are "transparent" to) questions about p itself (Moran, 2001, p. 67). The idea was made familiar by Gareth Evans, who writes: "I get myself in a position to answer the question whether I believe that p by putting into operation whatever procedure I have for answering the question whether p" (1982, p. 225).

By adopting the deliberative stance in arriving at her self-ascriptions and so conforming to the Transparency Condition, the subject is guided by a commitment to 'rational authority', or the authority of justifying, as opposed to explanatory, reasons in determining her beliefs, desires and intentions. Commitment to the Transparency Condition lies at the core of Moran's understanding of the link between self-knowledge and rational agency, as the subject exercises rational control over his mental life only to the extent that he undertakes the deliberative stance toward it. As Moran puts it, "the goal of deliberation, whether practical or theoretical, is conviction, about what to do or what to think" (2001, p. 131). Consequently, a non-evidential first person access to one's beliefs is a basic requirement of rational agency; a failure of transparency in one's deliberation amounts to a failure of self-knowledge, a failure to reach a fully conscious or first-personal state of knowledge of one's mental life.

Responsibility, reflection, and responsiveness to reasons

According to what I have referred to as the regulative model, self-knowledge is essential for maintaining rational coherence in one's

mental life. Given that it is in virtue of our capacity to exercise reflective control over our mental states that we can be held responsible for them, self-knowledge is also essential to rational agency. Not only can our second-order beliefs about the reasonableness of our first-order states serve as reasons for those states but, as far as our status as rational agents is concerned, they are the primary reasons that "rationally motivate" those states.[2] This is not to say that, on this view, rational belief formation must always involve second-order reflection on the soundness of the reasons for it. A subject's belief that *p* (say, that a mouse has taken up residence in her house) may be based on a first-order awareness of pieces of evidence—a hole chewed in a bag of rice, what appear to be mouse droppings on the shelf—that serve as reasons to motivate and warrant the belief that *p*. However, if she is to be held responsible for her first-order state, the subject must be capable of forming a judgment concerning whether or not it is justified through second-order reflection on the justificatory force of the first-order belief, as well as on the reasoning that supports it. Her focus is not on whether the hole in the bag and droppings were caused by a mouse, but rather on whether her evidentiary beliefs about these things warrant a belief in its presence. And, according to the model we are considering, this higher-order judgment must determine whether she holds the belief that a mouse is indeed in the house.

What does this involve? Suppose a subject believes that *p* for reasons *q* and *r*. First of all, if she is to reflect on her belief that *p* and her reasons for it she must know what that belief and those reasons are—she must form true second-order beliefs about them. She then deliberates on the soundness of the first-order belief by examining those beliefs that serve as reasons for it, as well as the reasoning that connects them to it. This includes judging whether they are justified, whether there are any fallacies in her reasoning, and whether the evidence represented by those beliefs is sufficient to support the belief they are taken to motivate. Having successfully applied her knowledge of epistemic norms to her reasons and reasoning, she may either (1) find that everything meets the epistemic mark, in which case she endorses the belief as one she ought to have and, so, maintains it, or (2) find some fault in her reasoning and judges that she ought not hold the belief, at which point she changes her mind. It is in this way that the subject assumes responsibility for her belief.

This picture faces some challenges. I shall mention two. First, let us suppose that such second-order judgment is possible. The proponent of reflective control claims that a subject's second-order judgment (to the effect that her first-order *prima facie* reasons and reasoning are in good order) is what ultimately motivates her belief that *p*. That is, what directly motivates a subject's first-order belief that *p* for which she may be held responsible are not her first-order judgments about the world but her second-order belief that the normative constraints on belief have been met. It is in light of her second-order judgment that the belief that *p* is sound and ought to be believed that she decides to believe it. But can these second-order judgments really play the motivational role envisioned for them?

David Owens points out that in order to reflect on the reasonableness of her belief that *p*, the subject must already have a first-order awareness of the evidence that prompts that belief. In exercising reflective control over her mental states, she engages in second-order judgment the purpose of which is to ensure her reasonability by explicitly acknowledging the evidentiary force of the reasons she already has. But what do the subject's higher-order judgments that she has those reasons for her belief, and that they suffice for the reasonableness of that belief, add to the motivational equation? How do they exert an independent rational influence on—count as reasons for adopting—her belief? As Owens puts it, "[i]f you already have a non-reflective awareness of the reasons which ought to motivate you, how does the judgment that you ought to be moved by them help to ensure that you are so moved? Such judgments", he concludes, "look like an idle wheel in our motivational economy…." (2000, p. 18). Such higher-order judgments seem only to confirm what is already accomplished by the available reasons for belief. The picture we are given is of a mind turned inward, inquiring into the rational standing of its own contents. It is unclear how the product of this inner inquiry—a second-order mental state that pronounces on the epistemic fitness of other mental states—is able to serve as the primary reason to adopt a first-order belief about the world.

A second objection concerns an infinite regress that potentially threatens the rational agency model. According to the model, warrant for, and the reliable truth of, a subject's second-order beliefs and judgments about the rational standing of her first-order mental states is required for the regulative role they play in the maintenance of her rationality. In short, if our second-order judgments did not rise to the level of *knowledge* of our first-order states, we could not ensure that our first-order states were rational. So, as Burge says,

> one must have an epistemic entitlement to one's judgements about one's attitudes. [Furthermore], if reflective judgements were not normally true, reflection could not add to the rational coherence or add a rational component to the reasonability of the whole process. It could not rationally control and guide the attitudes being reflected upon…. (1998, pp. 249–250)

On the one hand, the epistemic warrant for and security of self-knowledge is required for critical rationality, which itself is needed to regulate one's first-order mental life. The reflective second-order judgments we arrive at regarding our first-order mental states could only fulfill their regulative role if they were largely true and if we were entitled to them. That is, if the second-order judgment that one ought, or ought not, believe that *p* was not one to which were entitled, it could not serve as a reason to form, maintain, or discard the belief that *p*. So, second-order judgments must count as knowledge in order to play their regulative role. But what explains the fact that our second-order judgments about our first-order states are judgments to which we are entitled? That our first-order states are in accord with reason is explained by the regulative activity of our second-order judgments. But what explains how those second-order judgments are normally sound? That they must be is required for the role they play. However, to make this point is not to explain how they acquire this status. If the rationality of our second-order judgments were explained in the same way that the rationality of our first-order beliefs is explained, then a third-order of belief would be needed to regulate our second-order beliefs. But since our third-order beliefs can only perform that regulatory role if they are themselves rational (that is, if we are entitled to them and they are reliably true), the regulative account is set off on an infinite regress. If, however, our second-order judgments do not themselves require regulatory oversight to keep them in accord with reason, and they can (somehow) remain reasonable without higher-order supervision, then such supervision cannot be necessary for rationality *per se*. So the appeal to higher-order regulative intervention cannot be *required* to account for the rationality of our first-order states.

It is beyond the scope of the present discussion to offer a detailed alternative to the regulative model. But it may relieve potential anxieties about rationality to see that it can survive the loss of a second-order regulative level of mind. To that end, I will provide a sketch of how we might have both rationality and groundless authority about our own mental states without need for higher-order regulation. I begin with a sketch of a Wittgensteinian expressivist account of authoritative self-ascriptions.

Expressivism and authoritative self-ascription

Denying that second-order deliberation and the self-knowledge it presupposes must figure in the rational motivation of our mental states is consistent with an expressivist understanding of authoritative self-ascriptions. On this view, we need not engage in second-order cognition to authoritatively self-ascribe our mental states; rather, this capacity is explained by first-order linguistic expressive know-how, or the learned ability to express one's mental states using linguistic constructions (for example, the learned ability to utter "I'm I pain", in place of a moan). According to Wittgensteinian expressivism, the non-evidential basis and reliable truth of authoritative self-ascriptions are explained by the fact that such utterances ascribe the very beliefs they express.[3] The essential claim is that, contrary to superficial syntactic appearances, utterances of '*p*' and 'I believe (or desire, intend, etc.) that *p*' typically express the same mental state of belief (or desire, intention, etc.) that *p*. However, it remains that, as indicated by their differing truth conditions, they mean different things. For a special class of self-ascriptions—those that express the states they ascribe—meaning and expressive content diverge.

The account of the basic asymmetries distinctive of self-ascriptions of mental states is then this: If my utterance of 'I believe that *p*' serves to ascribe to me the belief that *p*, it follows that my *utterance* will be true if and only if I do in fact have that belief. This is merely a commonplace about the concept of truth. But according to the expressivist thesis, in making the utterance I also *express* the belief that *p*. This second claim is what defines any expressivist approach to the problem. Consequently, if I am sincere in my utterance of the self-ascription (i.e., I have the belief I express), then it follows that my utterance must be true. And this accounts for why, when I utter sincere self-ascriptions of my mental states, I will, saving exceptional cases like self-deception, "get them right".[4] That our self-ascriptive utterances can serve as expressions of first-order mental states—the same states they ascribe—rather than expressions of second-order beliefs about those first-order states, also explains why we can make them immediately and effortlessly, without appeal to evidence. As an expression of pain, an utterance of "That hurts!" is no more in need of justificatory evidence than an exclamation of "Ouch!" The immediacy of the self-ascription of pain follows from understanding "That hurts!" as a learned, substitute expression of pain itself, and not a higher-order expression of belief about one's being in such a state.[5]

To see the epistemically deflationary character of self-ascriptions earned by the expressivist view, consider the analogous case of explicit performatives. My *saying* 'I promise to do x' will (in the appropriate context) bring it about that I do promise to do x and, so, provides a guarantee that what I say is true (see Sinnott-Armstrong, 1994). My promissory utterance is assured of truth because my saying that I promise normally makes it that case that I do promise. The guarantee that what I say is true is limited to the first person present tense since neither *your* saying that I promise, nor my saying that I *did* promise, can bring it about that I do promise. But my promissory utterance is not an assertion—it does not express a belief (e.g., to the effect that I promise to do x) and, so, it does not express any knowledge that I have. There is no epistemological accomplishment involved in my getting it right, and no special powers of detection underwrite the assurance that my self-ascription of my promise is true.

Analogously, when (as expressivists claim) first-person self-ascriptive utterances express the very states that they ascribe, then they express pains, desires, or fears, not second-order beliefs about those states. *A fortiori*, such self-ascriptions do not express justified true (second order) beliefs about our mental states—they do not express *knowledge*. Nonetheless, our self-ascriptions are guaranteed to be true whenever they are sincere, since the conditions required for their truth and the conditions required for their sincerity are the same (namely, that the self-ascriber has the mental state she both ascribes and expresses). In such cases, we need not express any second-order beliefs about our mental states and, once again, the reliable truth of our self-ascriptions is not underwritten by any special powers of detection. If knowledge is justified true belief about that topic, then knowledge is not on the table.

For our present purposes, the important feature of this account is that it shows us how to make sense of the distinctive asymmetries between first-person and other-person ascriptions of mental states without making any use of second-order thoughts. This is not, of course, to deny that we can and sometimes do have such thoughts: we sometimes (say, in a therapeutic setting) adopt what Moran calls the theoretical stance towards ourselves, and the upshot might be a self-ascriptive utterance that expresses a second-order belief about a first-order mental state. But from that stance—the same stance others might take toward us—we cannot speak with immediacy or authority about ourselves.[6] Just as our ability to reliably self-ascribe mental states might be only a first-order accomplishment, so might our ability to maintain a rationally coherent mental life.

Language, rationality, and the mental

According to Donald Davidson, rationality, thought, and speech are interdependent phenomena, the understanding of which requires that we focus on the communicative situation and ask what is needed for a hearer to successfully interpret the words of a speaker.

Two interrelated ideas, both drawn from this theory of interpretation, inform his view of rationality: the holism of the mental and the Principle of Charity. According to the former, a single belief, desire, or intention that *p* depends for its identity on the relations it bears to a host of other propositional attitudes. As he summarizes it with respect to beliefs,

> Because of the fact that beliefs are individuated and identified by their relations to other beliefs, one must have a large number of beliefs if one is to have any. Beliefs support one another and give each other content. Beliefs also have logical relations to one another. As a result, unless one's beliefs are roughly consistent with each other, there is no identifying the contents of beliefs. A degree of rationality or consistency is therefore a condition for having beliefs. (Davidson, 2001, p. 124)

Given that every other propositional attitude depends for its identity on a great many beliefs, the point extends a wide range of mental states. As Davidson writes: "[t]here are … no beliefs without many related beliefs, no beliefs without desires, no desires without beliefs, no intentions without both beliefs and desires" (2001, p. 126).

Contrast this with the regulative model of rationality, according to which failures of rationality amount to failures to effectively monitor and control one's first-order states through second-order scrutiny of them. If we fail and so lapse into irrationality, we would still have all our first-order states, however irrational they will have become. But on Davidson's holistic model, that would not be possible; sufficient disarray would preclude the possibility of assigning beliefs, desires, and intentions to persons. The regulative model appears to presuppose a problematically high degree of atomism for our mental states.

Given the interconnected nature of mental states, an interpreter makes her way into the speech and thoughts of another holistically, as opposed to atomistically—it is a continuous

process whereby light dawns gradually and, over time, more fully on the whole. Throughout the process the interpreter must deploy the Principle of Charity. This principle

> calls on us to fit our own propositions (or our own sentences) to the other person's words and attitudes in such a way as to render their speech and other behaviour intelligible. This necessarily requires us to see others as much like ourselves in point of overall coherence and correctness—that we see them as more or less rational creatures mentally inhabiting a world much like our own. (Davidson, 2004a, p. 35)

Although Davidson conceives of rationality in terms of logical consistency and coherence, his claims about charity and the holism of the mental can be sustained under other conceptions of what rationality requires:

> The issue is not whether we all agree on exactly what the norms of rationality are; the point is rather that we all have such norms, and that we cannot recognize as thought phenomena that are too far out of line. Better say: what is too far out of line is not thought. It is only when we can see a creature (or 'object') as largely rational by our own lights that we can intelligibly ascribe thoughts to it at all, or explain its behaviour by reference to its ends and convictions. (Davidson, 2004b, p. 97)

The assumption that those we seek to understand are rational by our standards must be in play from the outset—without it, the interpretive process could not get off the ground. As such, charity is not merely heuristic advice to the interpreter, to help choose between competing possible (in the sense of minimally plausible or reasonable) interpretations.[7] Without charity, there would be no grounds for differentiating a radically mistaken ascription of belief from a more plausible (reasonable) one given the same evidence (utterance plus behavior in a given surrounding). Insofar as "anything would go" in this regard, interpretation could not get started. Thus, the assumption of charity is a necessary condition for the possibility of interpretation, and so for a creature's counting as having speech and thought. As Davidson says, "[t]he policy of rational accommodation or charity in interpretation is not a policy in the sense of being one among many possible successful policies. It is the only policy available if we want to understand other people". He continues, "[w]e should not think of this as some sort of lucky accident, but as something built into the concepts of belief, desire, and meaning" (2004b, p. 36). In a sense, the communicative situation, in which charity is not optional, imposes rationality on us – the possibility of communication, and of finding others as having mental lives at all, depends upon each communicator's interpreting her interlocutor's utterances in such a way that they conform to her norms of rationality which, if they succeed in communicating, they must share.

According to the regulative model, a fully rational subject regulates her mental life and maintains intelligibility through second-order deliberation on the rational standing of ontologically distinct first-order states—rationality is a function of the subject's capacity to deploy her knowledge of rational norms in the analysis those states, through which she is able to exercise rational self-control. This suggests a compartmentalization of the mental that the holism and charity involved in interpretation eschews. Davidson thus offers what, in contrast with the regulatory model, might be called a "bottom-up" conception of rationality. Rationality is not imposed from the top down; rather, it is built into our propositional attitudes from the outset. As such, it is primarily a first-order affair.[8]

Conclusion

The foregoing can only be a starting point for the discussion of how to conceive of rationality without self-knowledge. Even so, any worries that rationality that might collapse with the loss of substantive self-knowledge and higher-order control presupposed by the regulative model of self-knowledge would be premature. At the very least we should recognize that the regulative model of mind is not mandatory, and that both rationality and authoritative self-ascription (traditionally understood as self-*knowledge*) might be understood without any essential involvement of higher-order thoughts and judgments.

I have argued that our supposed second-order judgments about the content and rational standing of our first-order states by which we are said to exercise control over those states could not do the job assigned to them. Since such second-order belief could not serve as a reason to adopt a first-order state, it would be, as Owens puts it, an idle wheel when it comes to the rational motivation of that state. But neither should we expect that we need second-order beliefs to play that role—our deliberations about what we ought to believe, desire, and intend can be guided by our understanding of the first-order reasons for them. And this is consistent with an account of authoritative self-ascription that denies that we have the kind of self-knowledge thought, by the advocate of the regulative model, to be necessary for rationality. If we can do without the regulatory model and its associated views of rationality and self-knowledge, perhaps we can erase yet another trace of the picture of mind as a theater made familiar by Descartes, Locke, and Hume. Even without an attentive audience, the play goes on.

Data availability

Data sharing not applicable to this article as no datasets were generated or analyzed during this study.

Received: 6 May 2020; Accepted: 26 August 2020;
Published online: 16 September 2020

Notes

1 Burge (Burge, 1998, p. 169) acknowledges the limited application of his analysis and has subsequently offered a quite different explanation—to be discussed below—of the knowledgeable status of judgments that are not self-referential.

2 The term 'rationally motivate' is borrowed from David Owens (2000). It is meant to "register the fact that reasons for belief produce belief … by explaining their product in a way that makes sense of it" (Owens, 2000, p. 17).

3 See, e.g., D. Finkelstein (2003, 2010) and R. Jacobsen (1996, 2009a). Dorit Bar-On (2004) offers what she terms a neo-expressivist understanding of self-knowledge. It is *neo*-expressivist in part because it preserves the idea that self-ascriptions express knowledge claims about mental states, where such knowledge is still understood on the model of justified true second-order beliefs about first-order mental states. The Wittgensteinian view I describe here sees no need to recover a traditional epistemology for our authoritative self-ascriptions. Furthermore, it avoids unnecessary complexity found in the neo-expressivist view motivated by a desire to preserve a traditional epistemological reading that coheres with the regulative model critiqued above.

4 See Jacobsen (2009b). The role of sincerity in the account is critical, but sometimes overlooked.

5 See Wittgenstein (Wittgenstein, 1963, p. 89, Remark 244) for the canonical expression of this idea.

6 It is worth noting that transparency of the sort Moran describes is a *consequence* of expressivism. Since expressivism says that my self-ascriptive utterances of 'I believe that *p*' typically express the first-order belief that *p*, it comes as no surprise that when asked for my grounds for my self-ascriptive utterance I will normally appeal to the grounds for that first-order belief. Again, deciding what to believe, desire, or intend is a "first-order affair". We employ the norms of rationality by which we form and maintain a rationally coherent mental life at the ground level, including the process by which we make the sort of adjustments the regulative model suggests requires a second-order monitoring function.See Finkelstein (2012) for more on the connection

between Moran's discussion of transparency and the expressivist understanding of self-ascriptions.

7 See Ramberg, 1989, pp. 71–77, for a detailed discussion of this matter.

8 Davidson does not offer an explicitly expressivist account of self-knowledge. Nonetheless, strong similarities may be found between Davidson's view of indirect discourse and the expressivist take on the grammar of self-ascriptions explained in §4 above. See. Davidson, 1984, pp. 106–107 and Jacobsen (2009a).

References

Bar-On D (2004) Speaking my mind. Oxford University Press, Oxford

Boghossian P (1998) Content and self-knowledge. In: Ludlow P, Martin N (eds) Externalism and self-knowledge. CSLI Publications, Stanford, Originally in: (1989) Philosophical Topics 17:5–26, pp 149–173

Burge T (1998) Our Entitlement to self-knowledge. In: Ludlow P, Martin N (eds) Externalism and self-knowledge. CSLI Publications, Stanford, Originally in: Proceedings of the Aristotelian Society, New Series 96:91–116, pp 239–263

Davidson D (1984) On saying that. Truth and interpretation. Oxford University Press, Oxford, Originally in: Synthese 19:130–146

Davidson D (2001) The emergence of thought. Subjective, Intersubjective, Objective. Oxford University Press, Oxford, Originally in: (Sept. 1999) Erkenntnis 51(1):511–521

Davidson D (2004a) Expressing evaluations. Problems of rationality. Oxford University Press, Oxford, Originally published as: Lindley Lecture, University of Kansas (1984)

Davidson D (2004b) Representation and interpretation. Problems of rationality. Oxford University Press, Oxford, Originally in: (1990) W.H. Newton-Smith and K.V. Wilkes (eds) Modeling the mind. Oxford University Press, Oxford, pp 13–26

Evans G (1982) The varieties of reference. Oxford University Press, Oxford

Finkelstein D (2003) Expression and the inner. Harvard University Press, Cambridge

Finkelstein D (2010) Expression and avowal. In: Jolley K (Ed.) Wittgenstein: key concepts. Acumen, Durham, pp 185–198

Finkelstein D (2012) From transparency to expressivism. In:Abel G, Conant J (eds) Rethinking epistemology Vol. 2. De Gruyter, Berlin, pp 101–118

Jacobsen R (1996) Wittgenstein on self-knowledge and self-expression. Philos Quart 46(182):12–30

Jacobsen R (2009a) Davidson and first-person authority: parataxis and self-expression. Pacific Philos Quart 90:251–266

Jacobsen R (2009b) The duck quacks back: a reply to A. Minh Nguyen. Dialogue. Can Philos Rev 48(3):655–663

Moran R (2001) Authority and estrangement: an essay on self-knowledge. Princeton University Press, Princeton

Owens D (2000) Rationality without freedom. Routledge, London

Ramberg B (1989) Donald Davidson's philosophy of language: an introduction. Basil Blackwell, Oxford

Shoemaker S (1996) On knowing one's own mind. The first-person perspective and other essays. Cambridge University Press, Cambridge

Sinnott-Armstrong W (1994) The truth of performatives. Int J Philos Stud 2:99–107

Wittgenstein L (1963) Philosophical investigations. In: Anscombe GEM, Rhees R (eds) G.E.M. Anscombe (trans). Blackwell Publishers, Oxford

Acknowledgements

For comments on earlier drafts I am especially thankful to Rockney Jacobsen, as well as audiences at the following conferences: Canadian Philosophical Association Annual Congress (Montreal); Northwest Philosophy Conference (Salem); Self and Others in Wittgenstein and Contemporary Analytic Philosophy (Southhampton).

Competing interests

The author declares no competing interests.

Additional information

Correspondence and requests for materials should be addressed to S.B.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.